УДК 004.89

ПОДХОД К СОЗДАНИЮ СИСТЕМЫ СБОРА И ОБРАБОТКИ НОВОСТНЫХ СООБЩЕНИЙ ИЗ TELEGRAM-КАНАЛОВ

Остраухов Е.В. (ВКА), Андрушкевич Д.В. (ВКА)
Научный руководитель – кандидат технических наук, Андрушкевич Д.В.
(Военно-космическая академия имени А.Ф.Можайского)

Введение. В условиях стремительного увеличения объема данных, публикуемых в открытых источниках, автоматизация процессов сбора и анализа информации становится особенно актуальной. Telegram-каналы, как один из наиболее популярных источников оперативных данных, предоставляют широкий спектр сведений, включая новости, экспертные оценки и комментарии пользователей. Однако для эффективной обработки и фильтрации таких данных необходимы специализированные инструменты [1].

Цель данного исследования заключается в проведении сравнительного анализа различных методов сбора данных из Telegram-каналов, чтобы выявить наиболее эффективный подход к созданию системы сбора и обработки новостных сообщений значительного числа каналов (порядка 500). В работе рассматриваются три метода: использование Telegram API, web-парсинг и применение сети Tor. В результате исследования определяется подход, обеспечивающий максимальную скорость, точность и стабильность работы системы [2].

Основная часть. В данной работе рассмотрены три основных метода, которые могут быть применены для сбора данных из Telegram-каналов, и определено, какой из них позволяет наиболее эффективно обрабатывать большое количество каналов.

1. Telegram API – обеспечивает доступ к открытым каналам, позволяя получать сообщения без необходимости эмуляции веб-браузера. Для работы используется библиотека Telethon, однако этот метод требует авторизации и подвержен ограничению на количество запросов [3].
2. Web-парсинг – метод имитации пользовательского взаимодействия с веб-версией Telegram с помощью Selenium и BeautifulSoup. Этот метод позволяет извлекать данные без применения API, но зависит от изменений интерфейса платформы и требует значительных вычислительных ресурсов [4].
3. Сбор через Tor – метод анонимизированного поиска каналов и контента с использованием поисковых систем и каталогов через сеть Tor. Для реализации применяются библиотеки requests и BeautifulSoup, однако этот метод отличается низкой скоростью и нестабильностью результатов [5].

На основе проведенного анализа предлагается комбинированный подход, включающий применение Telegram API в сочетании с web-парсингом в качестве резервного метода. Такой гибридный метод позволит минимизировать риски, связанные с ограничениями API, и обеспечит стабильную и масштабируемую обработку данных.

После сбора сообщений предлагается осуществлять их фильтрацию по ключевым словам, временному диапазону и классификацию с применением методов машинного обучения. Классификатор, основанный на нейросетевой модели transformers, позволит определять тональность и тематику сообщений. Данный этап необходим для отсеивания нерелевантной информации и концентрации на наиболее значимых данных [6].

Выводы. Проведенное сравнение методов позволило определить, что наиболее эффективной стратегией для обработки большого количества информации собираемых из Telegram-каналов является комбинация API и web-парсинга. Данный подход позволит минимизировать ограничения каждого из методов в отдельности. Для классификации и фильтрации данных предлагается применение нейросетевых моделей transformers, что обеспечит точность классификации текстовой информации из социальных платформ. В

дальнейшем планируется оптимизация алгоритмов фильтрации и повышение точности классификации текстов. Итогом исследования станет разработка системы, способного обрабатывать большие объемы данных из Telegram-каналов, обеспечивая их классификацию и фильтрацию.

Список использованных источников:

1. Selenium WebDriver Documentation [Электронный ресурс]. URL: https://www.selenium.dev/documentation/webdriver/ (дата обращения: 30.01.2025).
2. Telegram API Documentation [Электронный ресурс]. URL: https://core.telegram.org/api (дата обращения: 01.02.2025).
3. Шумилина М. А., Коробко А. В. Разработка чат-бота на языке программирования Python в мессенджере Telegram // Научные известия. – 2022. – №. 28. – С. 47-54.
4. Цхошвили, Д. З. Примеры использования технологии парсинга / Д. З. Цхошвили, Н. А. Иванова // Актуальные вопросы в науке и практике : сборник статей по материалам IV Международной научнопрактической конференции : в 5 ч. Самара, 11 декабря 2017 года. – Самара : ООО «Дендра», 2017. – С. 135–138.
5. Ryan Mitchell Web Scraping with Python COLLECTING DATA FROM THE MODERN WEB / Ryan Mitchell – First Edition – California: O'Reilly Media, Inc., 2015 – С. 74–88.
6. Simon Munzert, Christian Rubba Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining / Simon Munzert, Christian Rubba – United Kingdom: Wiley, 2015 – С. 223–248.

Остраухов Е.В. (автор) Подпись

Андрушкевич Д.В. (автор) Подпись